Conceptual Mapping of User's Queries to Medical Subject Headings

Yuri L. Zieman, PhD and Howard L. Bleich, MD
Beth Israel Deaconess Medical Center,
Harvard Medical School, Boston

This paper describes a way to map users' queries to relevant Medical Subject Headings (MeSH terms) used by the National Library of Medicine to index the biomedical literature. The method, called SENSE (SEarch with New SEmantics), transforms words and phrases in the users' queries into primary conceptual components and compares these components with those of the MeSH vocabulary.

*Similar to the way in which most numbers can be split into numerical factors and expressed as their product – for example, 42 can be expressed as 2*21, 6*7, 3*14, 2*3*7, – so most medical concepts can be split into "semantic factors" and expressed as their juxtaposition. Note that if we split 42 into its primary factors, the breakdown is unique: 2*3*7. Similarly, when we split medical concepts into their "primary semantic factors" the breakdown is also unique. For example, the MeSH term 'renovascular hypertension' can be split morphologically into reno, vascular, hyper, and tension – morphemes that can then be translated into their primary semantic factors – kidney, blood vessel, high, and pressure. By "factoring" each MeSH term in this way, and by similarly factoring the user's query, we can match query to MeSH term by searching for combinations of common factors.*

Unlike UMLS and other methods that match at the level of words or phrases, SENSE matches at the level of concepts; in this way, a wide variety of words and phrases that have the same meaning produce the same match. Now used in PaperChase, the method is surprisingly powerful in matching users' queries to Medical Subject Headings.

INTRODUCTION

In the early days of online searching of the MEDLINE database,[1] librarians were expected to type in (correctly spelled) each Medical Subject Heading (MeSH term). In 1981, PaperChase permuted the MeSH terms online, and shortly thereafter the National Library of Medicine began to provide permuted terms on the MeSH tape.[2] In 1983, Horowitz and his colleagues described a method of pointing users from title words to MeSH terms.[3] Shortly thereafter, MiniMedline[4] and PaperChase stored back pointers online, and now many providers of the MEDLINE database employ one or more methods of mapping users' queries to MeSH terms.

When users type a single word, if that word exists as a text word in the MEDLINE database, current mapping techniques work reasonably well. But if the user types a phrase that is neither a permuted MeSH term nor a pointer to one, or if the user concatenates several phrases in a natural expression, most mapping techniques perform poorly or, more often, fail completely.

The idea of representing a phrase by its concepts was influenced by classical works in linguistics. Naom Chomsky discussed how, on the basis of "surface structure" (the actual phrase) the deep structure (meaning) can be derived.[5] Igor Mel'cuk described a more detailed study of the relationship between text and meaning.[6,7] More recently, several authors have described interesting morphosemantic ways to extract meaning from a phrase,[8-11] but to our knowledge none of these approaches has led to the development of a practical working tool.

The best collection of medical language knowledge sources is contained in the Unified Medical Language System (UMLS) developed by the National Library of Medicine.[12] UMLS can be used to map queries to MeSH terms provided that the software can match the user's phrase to an entry that has a MeSH equivalent in one of the authority files. The goal of SENSE is to map multiword queries without the labor needed to create the authority files, the metathesaurus, the specialist lexicon, and the semantic network. To help achieve this goal, we have narrowed the problem in the following ways: First, we deal only with the medical domain. Second, we work only with short

1091—8280/97/$5.00

phrases – the kind of inputs that are reasonable for a user to type to query a bibliographic database. Third, we pick only the main subject of the query – the part that is most important for bibliographic retrieval. Finally, as in the case of numerical factors, we ignore (when doing so would not affect meaning) the order of appearance.

METHODS

To translate text into concepts, we use a notation that employs primary concepts that we call "semantic factors." There are two requirements for these semantic factors: 1) each factor should be a prime – that is, the concept that it represents should be so simple that we wouldn't want to split it further; and 2) we should be able to describe any concept in the medical domain, regardless of its complexity, by using the appropriate combination of semantic factors. Semantic factors are the "bricks" out of which any concept in medicine can be built. Examples of semantic factors are "heart," "pain," "walk," "bacteria," "arm," "high," "paralysis," "deficit," and "inflammation."

Just as prime numbers become scarcer and scarcer the higher we go, so semantic factors become scarcer and scarcer the more new concepts we factor. As of today, 3400 semantic factors plus 2700 proper names ("Alzheimer," "Hodgkin," "Wassermann," etc.), have sufficed. We don't apply factoring to drug names. Instead, we look up trade names in a separate dictionary and, when possible, use the generic name to point to the MeSH term.

Any discussion of concepts must deal with the relationships between them. When concepts are broken down into semantic factors, how can we measure how close one concept is to another? To put it another way, when a user types a query, we want to offer not only the MeSH term that most closely resembles that query, but also MeSH terms from concepts that are closely related to it.

Two models can be given as examples of measuring how close one concept is to another in factor representation. The first – a bit mathematical – represents each concept by a vector in multidimensional space, where each coordinate is a one or a zero corresponding to whether the factor contributes or does not contribute to the concept. The closeness of two concepts can be evaluated by the angle between their vectors.[8]

As an example, if two concepts each had two factors, we would have 4 cells with, say, (1,1) in the upper left and (0,0) in the lower right. These two concepts, which share no factors, would be 180 degrees apart, and thus totally unrelated. In contrast, (1,1) and (1,0), would be 90 degrees apart, and thus 50 percent related.

Since SENSE deals with 3400 factors (plus proper names and drugs), use of this method would imply a multidimensional space with over 6000 dimensions. For computational reasons, we use a simpler method to measure the closeness of relationships: we count the number of factors in common. As an example, let us consider the disease SCLEROTIC MYOCARDITIS. This phrase can be divided into its four primary semantic factors:

1 - INDURATION
2 - MUSCLE
3 - HEART
4 - INFLAMMATION

Let us denote the equivalence between the semantic factors and the name of the disease as

1,2,3,4 = SCLEROTIC MYOCARDITIS

Listed below are all combinations of three and two semantic factors (subsets of the four listed above), together with one example of the word or phrase that each encodes:

1,2,3 = MYOCARDIAL SCLEROSIS
1,2,4 = SCLEROTIC MYOSITIS
1,3,4 = SCLEROTIC CARDITIS
2,3,4 = MYOCARDITIS
1,2 = MUSCLE INDURATION
1,3 = CARDIAC SCLEROSIS
1,4 = INFLAMMATORY SCLEROSIS
2,3 = MYOCARDIUM
2,4 = MYOSITIS
3,4 = CARDITIS

SENSE treats these entities, in most cases without regard to what words or phrases are used to express them, as related, either wholly or in part, to SCLEROTIC MYOCARDITIS.

Structure of the Program

SENSE consists of three major components – a Medical Language Knowledge Base, a Semantic Analyzer, and a MeSH Semantic Index.

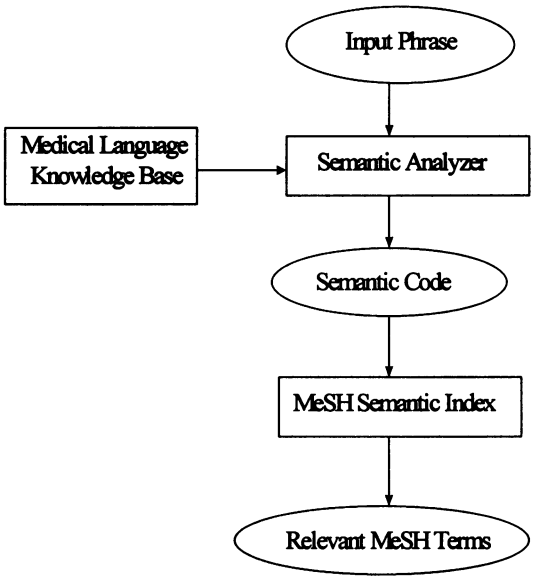
The Medical Language Knowledge Base describes how concepts are expressed in medical language. All descriptions in the Medical Language Knowledge Base are done at the level of quasimorphemes – small pieces of language having their own meaning. We say 'quasi' because for various practical reasons, some pieces consist of more than one morpheme.

The Semantic Analyzer, in conjunction with the Medical Language Knowledge Base, replaces what the user types with its prime semantic factors. We call the semantic factors produced for a particular word or phrase its "semantic code." The semantic code expresses the meaning of the phrase in our special notation.

Because the Semantic Analyzer deals with British as well as American spellings, singular as well as plural forms, and some common misspellings, there is no need for a separate dictionary to deal with these lexical variants. As of today, including all the variants and proper names, there are 9500 quasimorphemes in the Medical Language Knowledge Base. Ideally, using these quasimorphemes, the Semantic Analyzer would generate the identical, or nearly identical, semantic factors for all input phrases that have the same, or similar, meanings.

The third major component is the MeSH Semantic Index. This includes the semantic codes, not only for most MeSH terms, but for back pointers as well. Since we ignore order of presentation (the exceptions need not be discussed here), there is no need to permute MeSH terms or their back pointers.

The process of mapping an input phrase to a MeSH term may be graphically described as follows:



DISCUSSION

When users of PaperChase type medical phrases, the program negotiates with them to try to find the best relevant MeSH terms. In addition, the program stores what the user has typed, what SENSE has offered as possible matches, and what MeSH terms the user has selected. In this way, SENSE is continually tested and refined.

The examples listed below illustrate input phrases followed by MeSH terms suggested by SENSE.

HIGH PRESSURE IN THE LEFT RENAL ARTERY
- *hypertension, renovascular
- *hypertension, renal
- *renal artery
- *hypertension
- *arteries
- *nephrotic syndrome
- *nephrosis
- *kidney
- *pressure

SURGICAL REMOVAL OF A PORTION OF ONE BREAST
- *mastectomy, segmental
- *mastectomy
- *mammae
- *breast
- *fractionation
- *surgery, operative
- *surgery

MEDICATION USED TO TREAT HIGH CHOLESTEROL
- *anticholesteremic agents
- *drug therapy
- *hypercholesterolemia
- *cholesterol
- *drugs
- *bile
- *medicine
- *sterols

CANCER ORIGINATING IN THE BRONCHIAL TREE
- *carcinoma, bronchogenic
- *bronchial neoplasms
- *carcinoma
- *carcinogens
- *neoplasms
- *bronchi
- *wounds and injuries
- *trees

NURSES ASSISTING IN CHILDBIRTH
- *neonatal nursing
- *nurses' aids
- *labor
- *child
- *nurses
- *nursing

EKGs DONE IN OUTPATIENT CLINICS
- *electrocardiography, ambulatory
- *ambulatory care facilities
- *electrocardiography
- *endocardium
- *inpatients
- *outpatients
- *electricity
- *heart
- *patients
- *writing

FACE PAIN ATTRIBUTED TO THE FIFTH CRANIAL NERVE
- *trigeminal neuralgia
- *trigeminal nerve
- *facial neuralgia
- *facial nerve
- *facial pain
- *neuralgia
- *cranial nerves
- *face
- *pain
- *skull

IMMUNOLOGIC REJECTION OF ORGAN GRAFTS
- *transplantation immunology
- *graft rejection
- *organ transplantation
- *allergy and immunology
- *transplantation
- *immunity
- *science